基于车联网大数据的车险费率因子分析

高光远[1]　孟生旺[1,2]

(1. 中国人民大学应用统计科学研究中心,北京 100872; 2. 兰州财经大学统计学院,甘肃 兰州 730020)

[摘　要]　随着车辆网技术的不断成熟,车联网数据的应用价值日渐凸显。车联网大数据中包含着丰富的驾驶行为信息,这些信息对于改进传统的汽车保险定价模型具有重要的应用价值。如何从车联网大数据中提取出具有实际应用价值的信息,尚需进行大量细致的研究工作。本文基于车联网记录的速度-加速度数据,应用核密度估计和主成分分析,提取了一个驾驶行为因子,并在泊松分布假设下建立了索赔频率的广义可加模型。实证研究结果表明,本文提取的驾驶行为因子对被保险车辆的索赔频率具有十分显著的非线性影响,为汽车保险定价提供了一个新的费率因子,有助于进一步提高汽车保险定价结果的准确性和合理性。

[关键词]　车联网; 大数据; 汽车保险; 费率因子; 驾驶行为; 索赔频率

[中图分类号] F222.3　[文献标识码] A　[文章编号]1004-3306(2018)01-0090-11

DOI: 10.13497/j.cnki.is.2018.01.008

一、引　言

在财产保险公司,汽车保险的业务收入往往占到整个公司的 70% 左右,对公司的长远发展具有举足轻重的影响。在传统的汽车保险中,通常使用驾驶人和被保险车辆的一些先验信息对汽车保险进行定价,这些先验信息也称作费率因子。最常采用的费率因子包括驾驶人年龄、性别、驾龄,以及车辆的使用性质、座位数或吨位数等(孟生旺,2011)。随着车联网技术的逐步成熟和自动驾驶时代的来临,基于驾驶行为信息的定价方法受到越来越多的关注。驾驶行为信息是指汽车行驶的时间、路线、速度和加速度等信息,这些信息不同于传统的先验费率因子。先验费率因子是在投保时就可以获得的风险信息,虽然它们对驾驶人的未来风险具有一定影响,但先验费率因子与实际风险之间通常仅仅是一种间接的相关关系,而非因果关系。譬如,年轻男性在总体上的索赔频率高于其他驾驶人,但年轻男性与保险索赔之间并非具有直接的因果关系。年轻男性之所以具有较高的索赔频率,真正的原因可能是他们遇事容易急躁、喜欢开快车、驾驶时间较长、行驶范围较大等。也就是说,并非所有的年轻男性驾驶人都具有较高的索赔频率,而是那些驾驶行为不够谨慎的年轻男性才具有较高的索赔频率。因此,在传统的汽车保险中,使用先验费率因子为汽车保险定价,可能会使得一些风险较低的年轻男性驾驶人支付了较高的保险费,从而导致费率厘定结果的不公平性。

在传统的汽车保险定价中,广义线性模型的应用最为广泛,它是汽车保险定价的标准模型。在理论研究中,也有应用广义可加模型,广义线性混合模型,以及基于位置、尺度和形状参数的广义可加模型厘定汽车保险费率的应用案例(De Jong,2008; 孟生旺,2014)。最近几年,随机森林、神经网络、支持向量机等机器学习算法在汽车保险定价中的应用也受到了较多关注(Leo,2012; Liu,2014; 孟生旺,2017)。但是,这些方法仅仅

[基金项目]　本文获得教育部人文社会科学重点研究基地重大项目"基于大数据的精算统计模型与风险管理问题研究"(16JJD910001)、国家社科基金重大项目"巨灾保险的精算统计模型及其应用研究"(16ZDA052)以及中国人民大学 2017 年度"中央高校建设世界一流大学(学科)和特色发展引导专项资金"支持。

[作者简介]　**高光远**,中国人民大学统计学院风险管理与精算学博士后,研究方向:风险管理与精算; **孟生旺**,中国人民大学统计学院风险管理与精算学教授,兰州财经大学"飞天学者"讲座教授,研究方向:风险管理与精算。

是对定价模型的局部改进，并没有引入驾驶行为信息，都只是从先验费率因子中挖掘风险信息。

随着车联网技术的发展，保险公司能够采集到被保险车辆较为详细的驾驶行为数据，这为促进汽车保险定价的合理性和公平性创造了难得的机会。使用不同的设备采集到的驾驶行为数据略有区别，但通常包括GPS 经纬度、GPS 方向和速度、车辆感应器速度、引擎转速、三轴加速度等。这些数据通常以每秒或者以数秒为时间单位进行记录，可以详细地描述汽车每一个行程的时间、地点、轨迹、速度、加速度、方向和转弯等。在现有的理论研究和实际应用中，主要是基于汽车的行驶里程数为汽车保险定价，相应地，这种汽车保险被称作基于使用量的汽车保险(usage-based insurance，UBI) 。实证研究结果表明，行驶里程数对汽车保险的索赔频率具有十分显著的影响(张连增，2012; Ayuso，2016) ，又由于行驶里程数的采集和使用相对简单，所以在UBI 汽车保险的定价模型中，行驶里程数要么被作为定价基础使用，使得保费与行驶里程数成比例，要么被作为最主要的费率因子使用，使得保费在很大程度上依赖于行驶里程数。

除了行驶里程数以外，与风险相关的驾驶行为信息还包括汽车的行驶速度、加速度和驾驶时间等信息，关于这些驾驶行为信息在汽车保险定价中的应用研究还相对较少，且具体方法差异较大，譬如，Weinder (2016) 使用傅立叶分解法来区分不同的速度 - 加速度模式; Wüthrich(2017) 使用聚类分析对速度 - 加速度模式进行分类; Gao 和 Wüthrich(2017) 使用主成分分析和瓶颈神经网络提取速度 - 加速度模式。现有研究主要探讨了如何从高频车联网数据中提取可能的风险因子，并没有应用实际损失数据检验提取的驾驶行为因子是否对索赔频率具有显著影响。本文在 Gao 和 Wüthrich(2017) 研究成果的基础上，建立了预测索赔频率的泊松广义可加模型，并用实际损失数据进行了实证检验。交叉验证的结果表明，基于速度 - 加速度核密度估计的第二主成分对索赔频率具有十分显著的非线性影响，所以本文把该主成分定义为驾驶行为因子，该因子可以作为汽车保险定价的一个重要费率因子使用。

二、数据描述

本文使用的数据来自 1680 辆投保了交强险的汽车，数据集中既有车联网信息，也有交强险的索赔信息，还有驾驶人的年龄和性别信息。由于部分汽车续保了两年或三年，所以保单组合的总车年数为 2893，总索赔次数为 715，总平均的索赔频率为 24. 71% 。

本文选取了一周的车联网数据，该周一共记录到了 47045 次行程，平均每辆汽车有 28 次行程。一次行程定义为从发动机启动到发动机熄火。对于每次行程，车联网可以记录汽车在每一秒的速度、加速度、方向、转弯角度等信息。考虑到速度和加速度最能反映汽车的驾驶行为风险，所以本文重点研究如何从车联网记录的速度和加速度中提取用于汽车保险定价的费率因子。

对于汽车的行驶速度，本文使用 5km/h ~ 20km/h 范围内的车辆感应器速度。之所以选取这个速度区间进行分析，是因为本文仅对索赔频率进行建模，而绝大多数事故就发生在这个速度区间。不过，在对索赔强度进行建模时，就必须考虑汽车的高速行驶区间，因为少数重大事故常常发生在较高行驶的速度区间内。

汽车加速度的取值应该在 $-2\ m/s^2 \sim 2\ m/s^2$ 的范围内。为了剔除原始数据记录中的异常值，在数据分析时，把超出加速度正常范围的数据进行了截断处理，即将加速度小于 $-2\ m/s^2$ 的数值记作 $-2\ m/s^2$，将加速度大于 $2\ m/s^2$ 的数值记作 $2\ m/s^2$。

图 1 和图 2 显示了两辆代表性汽车的速度 - 加速度散点图，分别是数据集中的第 300 辆汽车和第 1100 辆汽车。可以看出，在速度相同的情况下，第 300 辆汽车有更多的急加速，尤其是在低速行驶的情况下，加速度超过 $1\ m/s^2$ 的比例远远大于第 1100 辆汽车。这两辆汽车的索赔数据显示，第 300 辆汽车在观察期发生过一次事故，而第 1100 辆汽车没有发生事故。在后面的建模分析中还可以进一步发现，对于相同的驾驶员年龄，第 300 辆汽车的驾驶行为更容易导致保险事故。这就意味着，汽车行驶的速度 - 加速度数据中确实隐含着重要的风险因子。由于车辆感应器的速度值是正整数，所以为了减少数据点的重叠，在图 1 和图 2 中，把速度进行了一个随机扰动，即把每个速度值都加上了一个来自(-0. 5，0. 5) 均匀分布的随机数，从而使得

散点图可以更加清晰地揭示速度和加速度的分布状况。

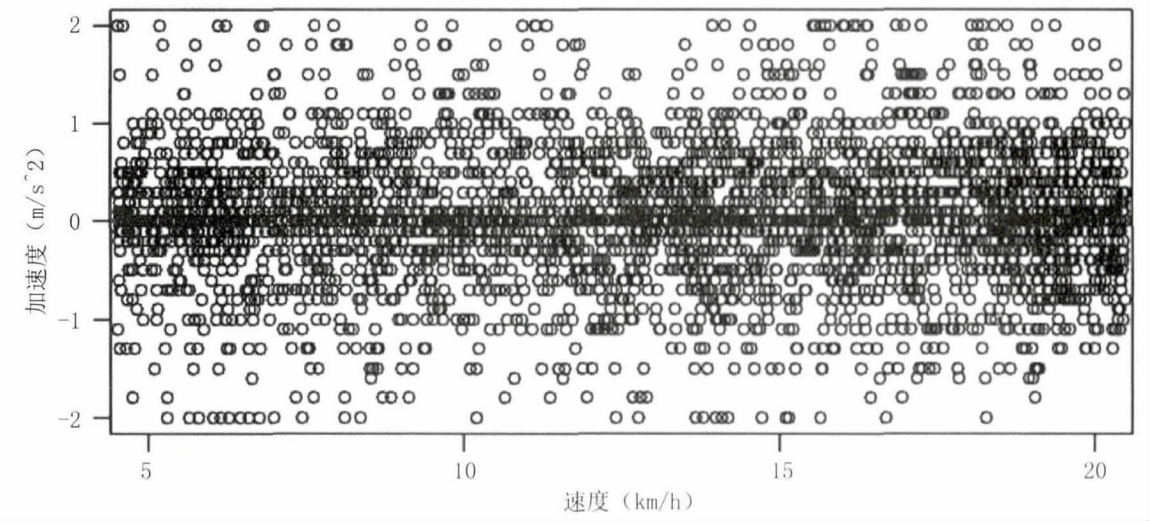

图 1　速度 - 加速度的散点图(第 300 辆汽车)

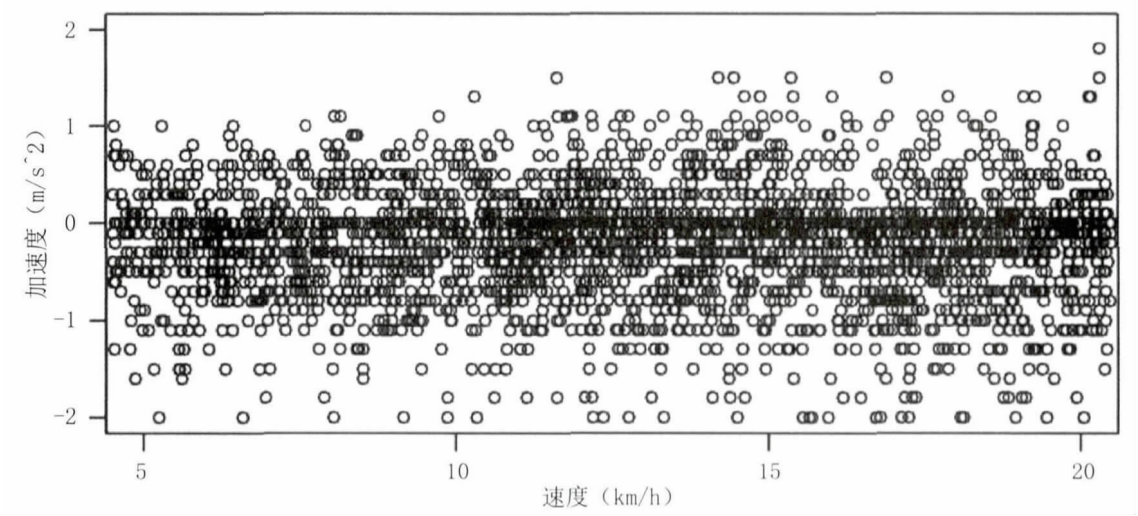

图 2　速度 - 加速度的散点图(第 1100 辆汽车)

三、速度 - 加速度的概率核密度估计

图 1 和图 2 虽然可以较为直观地揭示不同汽车在速度 - 加速度上的差异,但还很难将这种信息直接用于汽车保险的费率厘定。为此,把图 1 和图 2 中不同区域的点按其疏密程度进行量化,即估计该散点图的二维核密度。

散点图中速度和加速度的取值范围可以记为 $R = [5,20] \times [-2,2]$,在 R 上定义一个二维随机变量 $X = (v,a)^T \in R$,其中 v 和 a 分别表示汽车的速度和加速度。用 f_i 表示第 i 辆汽车在 R 上的速度 - 加速度的概率密度函数。

如果第 i 辆汽车的速度 - 加速度观测值为 $(x_1,\cdots,x_t) = ((v_1,a_1)^T,\cdots,(v_t,a_t)^T)$,则这辆汽车的速度 - 加速度的概率核密度估计可以表示为

$$\hat{f}_i(X) = \frac{1}{t}\sum_{s=1}^{t} K_H(X - x_s) \tag{1}$$

式(1) 中,K 为核函数,它是轴对称的二元密度函数,本文选取下述的二维高斯密度函数:

$$K_H(X) = 2\pi^{-1}|H|^{-1/2}\exp\left(-\frac{1}{2}X^T H^{-1/2}X\right) \quad (2)$$

H 为 2×2 的带宽矩阵，它是一个对称正定矩阵，其系数的作用类似于频率直方图中带宽的作用，H 决定了核密度函数的平滑程度。最优 H 可通过下面的式子求得

$$\hat{H} = \min_H E[\sum_{s=1}^{t}(\hat{f}_i(x_s; H) - f_i(x_s))^2]$$

其中等式右边的期望可通过交叉验证进行估计(Hastie，2009) 。

可以看出，某一点 x 的核密度估计为所有观测点对该点密度影响的平均值，譬如，观测点 x_s 对点 x 的密度影响为 $K_H(x - x_s)$ ，它随着 x_s 与 x 之间距离的增大而减小。换言之，某一点的核密度估计反应了所有观测点在该点的聚集程度。基于图 1 和图 2 的散点图求得的概率核密度估计如图 3 和图 4 所示。这两幅图均为灰度化的热力图，横轴表示速度，纵轴表示加速度。颜色越浅，表示该点上的概率越大。可以看出，第 1100 辆汽车的概率更多地聚集在加速度为零的附近，而第 300 辆汽车处于急加速区域的概率更大。

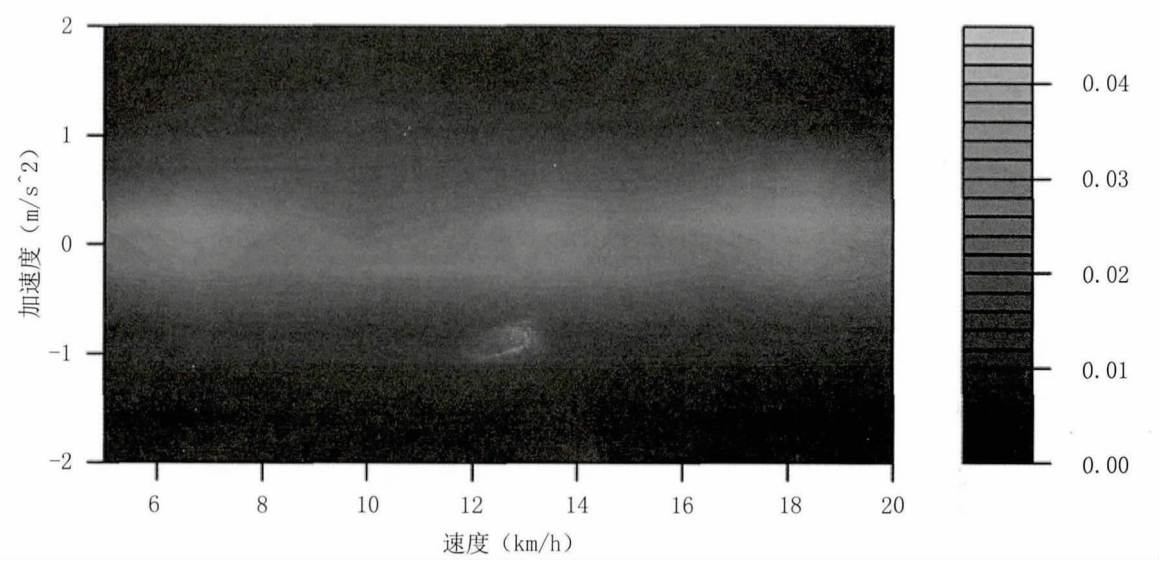

图 3　速度 - 加速度的核密度估计(第 300 辆汽车)

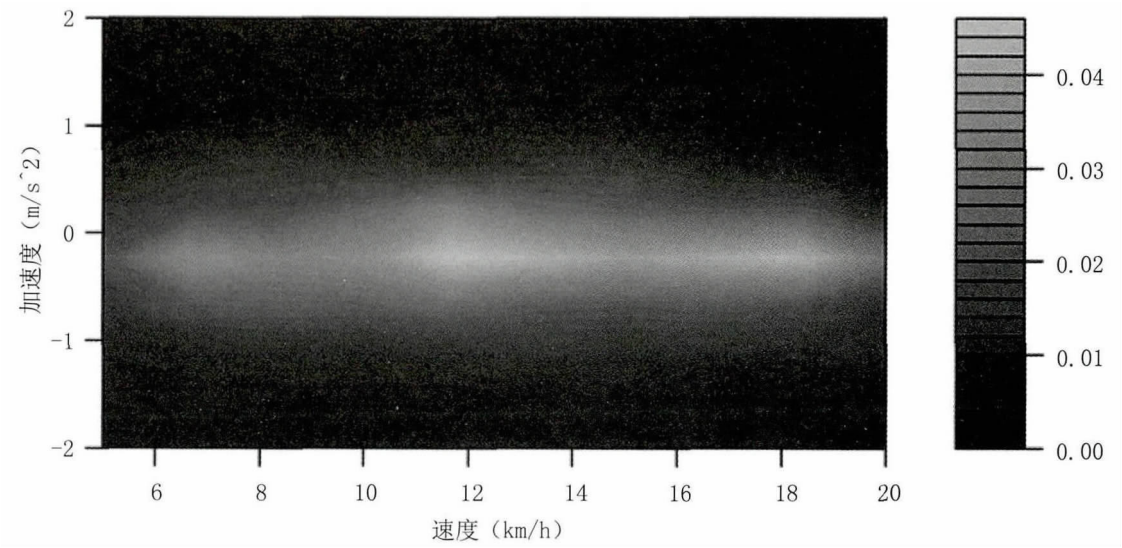

图 4　速度 - 加速度的核密度估计(第 1100 辆汽车)

四、驾驶行为因子

速度－加速度的核密度估计 $\hat{f}_i(X)$ 反应了第 i 辆汽车的驾驶行为。为了比较不同汽车的驾驶行为，从速度－加速度的核密度估计中选取固定的若干个格点来近似它的特征，如图 5 所示，该图在速度和加速度方向都选取了 10 个等间距的点，这些点对应的速度－加速度分别为 $x_1^0=(5.00,-2.00)^T$，$x_2^0=(6.67,-2.00)^T$，$\cdots$，$x_J^0=(20.00,-2.00)^T$，这些点上的核密度估计为 $[\hat{f}_i(x_j^0)]_{j=1:J}$，其中 $J=100$。把这些核密度估计进行标准化使得它们的总和为 1，即可得到下述的相对核密度估计：

$$z_{ij}=\frac{\hat{f}_i(x_j^0)}{\sum_{j=1}^{J}\hat{f}_i(x_j^0)} \tag{3}$$

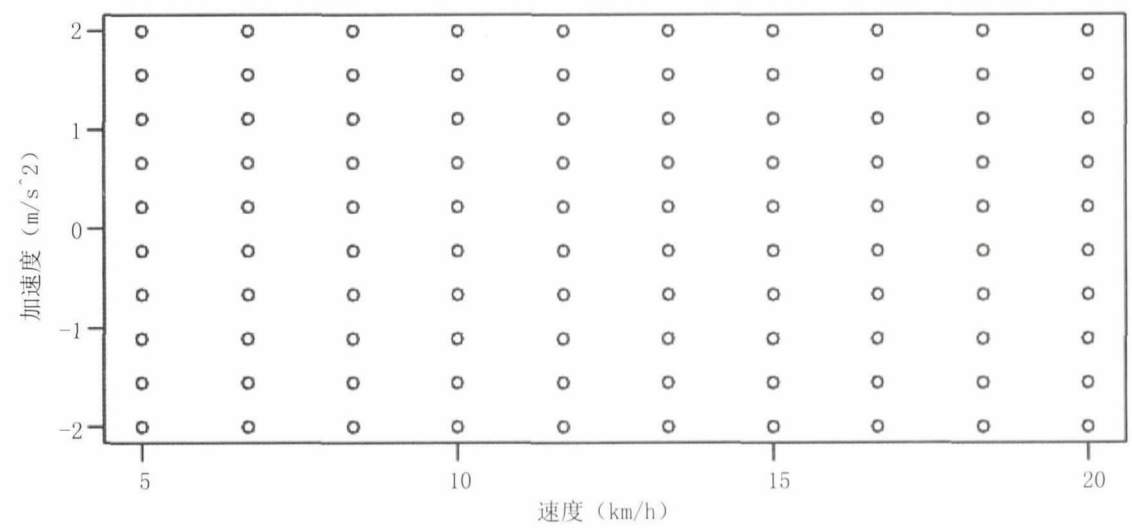

图 5　格点的选取

由此可见，第 i 辆汽车的驾驶行为可以用 J 维向量 $z_i=(z_{i1},z_{i2},\cdots,z_{iJ})^T\in\mathbb{R}^J$ 来刻画，相应地，所有汽车的驾驶行为可以用 $n\times J$ 的矩阵 $Z=(z_1^T,\cdots,z_n^T)^T$ 来刻画，该矩阵包含了 n 辆汽车的驾驶行为数据，其中第 i 行表示第 i 辆汽车的相对核密度估计，第 j 列表示所有汽车在 x_j^0 上的相对核密度估计。该矩阵可以称为驾驶行为协变量矩阵，包含了 100 个协变量。如果直接用这 100 个协变量建立预测模型，则很可能导致过拟合或发生共线性问题。下面通过对 Z 进行奇异值分解，并提取主成分来解决这个问题。奇异值分解的主要目的是从非方矩阵 Z 中提取和出险相关的主成分，对 Z 的秩没有要求。通常核密度估计矩阵 Z 为满秩矩阵，即有 J 个非零的奇异值。

假设矩阵 Z 的列向量之和为 $\mu=(\mu_1,\cdots,\mu_J)^T$，标准差为 $\sigma=(\sigma_1,\cdots,\sigma_J)^T$，则可以将矩阵 Z 标准化为

$$Z^0=(Z-Z_\mu)\times Z_{1/\sigma} \tag{4}$$

式(4) 中，Z_μ 是一个 $n\times J$ 矩阵，每一行的元素均为 μ^T；$Z_{1/\sigma}=\mathrm{diag}(1/\sigma_1,\cdots,1/\sigma_J)$ 是一个对角矩阵。经过标准化处理以后，矩阵 Z^0 的每一列的均值为 0，方差为 1。

在上述假设下，存在一个 $n\times J$ 的正交矩阵 U，一个 $J\times J$ 的正交矩阵 V，一个 $J\times J$ 的对角矩阵 $\Lambda=\mathrm{diag}(\lambda_1,\lambda_2,\cdots,\lambda_J)$，其中 $\lambda_1\geqslant\cdots\geqslant\lambda_J\geqslant 0$，可以将标准化以后的矩阵 Z^0 分解为(Hastie，2009)：

$$Z^0=U\Lambda V^T \tag{5}$$

其中 $\lambda_1\geqslant\cdots\geqslant\lambda_J\geqslant 0$ 称作矩阵 Z^0 的奇异值。V 的每一列称作荷载向量，其第 j 列记为 V_j。

对式(5) 的两边同时乘以 V，可以得到

$$Z^0V = U\Lambda \tag{6}$$

其中 Z^0V_j 称为矩阵 Z^0 的第 j 主成分。V_1 反映了在 J 维空间中方差最大的方向，V_2 反映了在 J 维空间中与 V_1 垂直的方差最大的方向，依次类推。

对于前文的速度－加速度数据，一共可以提取 100 个主成分，图 6 显示了前 10 个主成分可以解释的累积方差比例。可以看出，第一主成分和第二主成分解释的方差比例相对较大，而其他主成分所能解释的累积方差相对较小。在后面的预测模型中，把对索赔频率具有最显著影响的主成分定义为驾驶行为因子。

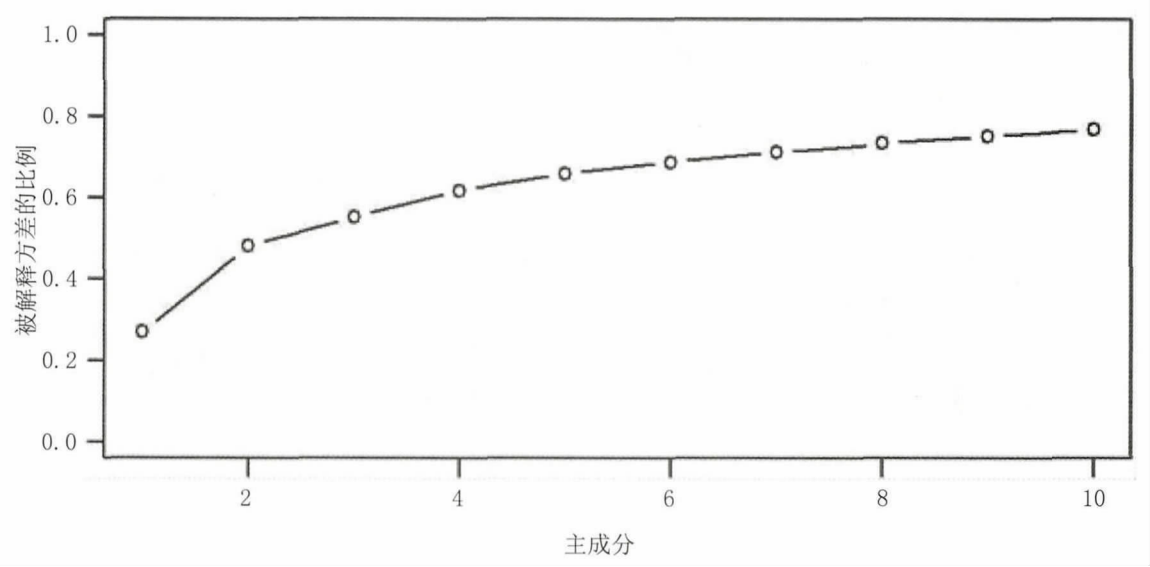

图 6　主成分所解释的累积方差

五、索赔频率的广义可加模型

下面讨论如何基于驾驶行为因子建立索赔频率的广义可加模型。假设每辆汽车的索赔次数服从泊松分布，泊松参数与保险期间的车年数成比例，且与驾驶人年龄和驾驶行为因子有关。考虑到驾驶人年龄和驾驶行为因子都是连续型变量，所以可以建立广义可加模型，并用自然三次样条曲线来拟合驾驶人年龄和驾驶行为因子对索赔频率的影响。关于广义可加模型的详细讨论可以参见 Wood(2006)。

在上述假设下，索赔频率的广义可加模型可以表示为

$$\begin{cases} Y_i \sim \text{Poisson}(\lambda_i e_i) \\ \lambda_i = \exp[s_1(age_i) + s_2(pz_i^m)] \end{cases} \tag{7}$$

式(7)中，Y_i 为第 i 辆汽车的索赔次数随机变量，服从参数为 $\lambda_i e_i$ 的泊松分布，其中 λ_i 表示第 i 辆汽车的索赔频率，e_i 表示第 i 辆汽车的风险暴露(车年数)；age_i 为第 i 辆汽车的驾驶人年龄；$pz_i^m = Z_i^0 V_m$ 为第 i 辆汽车的第 m 主成分；s_1 和 s_2 分别表示作用于驾驶人年龄和驾驶行为因子的自然三次样条曲线。

三次样条曲线是平滑连接的多段三次曲线，在连接的节点处要求曲线连续且二次可导。自然三次样条曲线还要求曲线在两端节点的外延部分为直线。自然三次样条曲线一般表示为基函数的线性组合。譬如，作用于驾驶人年龄的自然三次样条曲线 s_1 可以表示为:

$$s_1(age_i) = \sum_{k=1}^{K} \beta_k b_k(age_i) \tag{8}$$

其中$\{b_k: k = 1, \cdots, K\}$为基函数。本文使用满秩的自然三次样条曲线，它的基函数为多项式函数，节点放置在每个年龄值上。假设样本里有 K 个年龄值，则 s_1 有 K 个基函数，且有 K 个自由度。这是因为 K 个年龄值意味着 K 个节点，它们把一维空间划分为 K + 1 段。把这 K + 1 段自由度为 4 的三次曲线连接起来，并

保证在节点两次可导，则该曲线有 $4(K+1)-3K=K+4$ 个自由度，同时要求曲线在两端节点的外延部分为直线，所以最终的自由度为 $K+4-2\times2=K$。

基于前文的数据，使用不同的主成分和驾驶人年龄的组合，可以建立九种不同形式的模型，各种模型的结构如表 1 所示，其中 $pz_i^m(m=1,2,3,4)$ 表示第 i 辆汽车的第 m 主成分。

本文检验了前四个主成分对索赔频率的影响，分别对应模型 2、模型 3、模型 4 和模型 5。结果表明，第三主成分和第四主成分对索赔频率的影响不显著，所以在模型 6、模型 7、模型 8 和模型 9 中，仅仅考虑了第一主成分和第二主成分对索赔频率的影响。在模型 6 和模型 7 中，应用自然三次样条对第二主成分进行了平滑处理，而在模型 8 和模型 9 中，第二主成分被表示为线性函数的形式。

索赔频率的广义可加模型

表 1

模型编号	模型表达式	模型编号	模型表达式
1	$Y_i \sim Poisson(\lambda_i e_i),\lambda_i=\exp[s_1(age_i)]$	6	$Y_i \sim Poisson(\lambda_i e_i),\lambda_i=\exp[s_1(age_i)+s_2(pz_i^1)]$
2	$Y_i \sim Poisson(\lambda_i e_i),\lambda_i=\exp[s_2(pz_i^1)]$	7	$Y_i \sim Poisson(\lambda_i e_i),\lambda_i=\exp[s_1(age_i)+s_2(pz_i^2)]$
3	$Y_i \sim Poisson(\lambda_i e_i),\lambda_i=\exp[s_2(pz_i^2)]$	8	$Y_i \sim Poisson(\lambda_i e_i),\lambda_i=\exp[s_1(age_i)+pz_i^1]$
4	$Y_i \sim Poisson(\lambda_i e_i),\lambda_i=\exp[s_2(pz_i^3)]$	9	$Y_i \sim Poisson(\lambda_i e_i),\lambda_i=\exp[s_1(age_i)+pz_i^2]$
5	$Y_i \sim Poisson(\lambda_i e_i),\lambda_i=\exp[s_2(pz_i^4)]$		

为了对前述九种模型的拟合优度进行比较，可以计算广义可加模型的偏差。参数为 λ_i 泊松分布的概率函数为

$$Pr(Y_i=y_i)=\frac{\lambda_i^{y_i}e^{-\lambda_i}}{y_i!}$$

所以，泊松广义可加模型的偏差函数可以表示为：

$$D(y,\hat{\lambda})=2(l(y)-l(\hat{\lambda}))=2\left(\sum_{i=1}^{n}\hat{\lambda}_i e_i-y_i+y_i\log\frac{y_i}{\hat{\lambda}_i e_i}\right) \tag{9}$$

其中 $\hat{\lambda}$ 为参数 λ 的极大似然估计值，$l(y)$ 为饱和模型的对数似然函数值，$l(\hat{\lambda})$ 为所研究模型的对数似然函数值。当 $y_i=0$ 时，等式右边的第 i 项取 $\hat{\lambda}_i e_i$。

一般而言，偏差越小，表示模型对观察数据的拟合效果越好。

为了避免过拟合，下面应用十折交叉验证比较不同模型的偏差。首先把数据集 C 分为 10 个大小相近的互不重叠的子集，分别记作 $C_1,\cdots,C_{10}$，它们满足 $\bigcup_{l=1}^{10}C_l=C$。然后做 10 次模型拟合，其中第 l 次的模型拟合和偏差计算过程如下：

(1) 选取 $\Omega_{C_l}=\{(age_i,pz_i^m,y_i):i\in C_l\}$ 作为测试集，$\Omega_{-C_l}=\{(age_i,pz_i,y_i):i\notin C_l\}$ 作为训练集，并将参数估计结果记为 $\hat{\lambda}^{-C_l}$。

(2) 计算第 l 个模型的偏差

$$D_l=D((y_i)_{i\in C_l},\hat{\lambda}^{-C_l}) \tag{10}$$

最后计算前述 10 个模型的平均偏差：

$$\bar{D}=\frac{1}{10}\sum_{l=1}^{10}D_l \tag{11}$$

对于前述的九种泊松广义可加模型，交叉验证求得的平均偏差如表 2 的第 4 列所示，各个协变量的显著

性如表2的第3列所示。可以看出，驾驶人年龄和第二主成分对索赔频率的影响最为显著。模型1的唯一协变量为驾驶人年龄，模型3的唯一协变量是第二主成分，模型1的偏差小于模型2的偏差，说明第二主成分比驾驶人年龄对索赔频率的解释能力更强，所以将第二主成分定义为驾驶行为因子。

表2的第5列是基于全样本模型计算的校正后决定系数，它考虑了模型复杂度对决定系数的惩罚，所以也能在一定程度上防止过拟合。校正后的决定系数越大，表示模型的拟合效果越好。可以看出，无论采用哪个准则，在上述九个模型中，第7个模型是最优模型。

在模型7中，对驾驶人年龄和驾驶行为因子都采用自然三次样条进行了平滑处理。在模型9中，仅对驾驶人年龄采用自然三次样条进行了平滑处理，而驾驶行为因子被表示为线性函数的形式。比较模型7和9的偏差可以看出，模型7的偏差更小，说明驾驶行为因子对索赔频率的影响效应是非线性的。

模型比较

表2

模型编号	协变量	在5%水平下 协变量的显著性	交叉验证的平均偏差	校正后的 决定系数
1	驾驶人年龄	显著	0.8095622	0.00771
2	第一主成分	不显著	0.8070473	0.00299
3	第二主成分	显著	0.8034685	0.00786
4	第三主成分	不显著	0.8123352	(0.00196)
5	第四主成分	不显著	0.8123679	0.00015
6	驾驶人年龄，第一主成分	第一主成分不显著	0.8056216	0.01050
7	驾驶人年龄，第二主成分	显著	0.8026885	0.01450
8	驾驶人年龄，第一主成分	第一主成分不显著	0.8098052	0.00737
9	驾驶人年龄，第二主成分	显著	0.8039764	0.01120

最优模型(模型7)的参数估计结果如表3、表4所示，其中驾驶人年龄的有效自由度约为6，驾驶行为因子的有效自由度约为3，它们分别表示自然三次样条曲线 $\hat{s}_1$ 和 $\hat{s}_2$ 的自由度。从输出结果中的P值来看，驾驶人年龄和驾驶行为因子对索赔频率都具有显著的非线性效应。

最优模型(模型7)的输出结果

表3

	参数估计值	标准误	Z值	P值
截距	-1.42374	0.03828	-37.19	< 2e-16

平滑项的有效自由度和显著性检验

表4

	有效自由度(edf)	Chi. sq值	P值
s(age)	5.714	20.92	0.0036
s(pz2)	3.116	24.38	6.67e-05

图7和图8分别揭示了驾驶人年龄和驾驶行为因子对索赔频率的影响效应(对数尺度)，实线表示影响效应，上下两条虚线表示95%的置信区间。横轴上的每条竖线代表处于该驾驶人年龄或驾驶行为因子上的一辆汽车。对于车辆数较少的年龄段或者驾驶行为因子段，模型对该年龄段或者驾驶行为因子段的效应估

计存在较大的不确定性，在图中表现为较宽的置信区间，譬如，图 7 中驾驶人年龄大于 60 岁的部分，图 8 中驾驶行为因子小于 -20 的部分，置信区间都很大，表明模型在这个区间的估计结果存在较大的不稳定性。在实际应用中，为了获得相对稳定的费率厘定结果，可以把驾驶人年龄大于 60 岁的汽车段归为一类，把驾驶行为因子小于 -20 的汽车归为一类。

图 7 表明，驾驶人年龄的变化对索赔频率的影响是非线性的，在 30 岁以下的区间，索赔频率随着年龄的增加而上升，但由于该区间的车辆数较少，结果的可靠性较低；在 30 ~ 40 岁的区间，随着驾驶人年龄的增加，索赔频率略有下降，譬如，在给定驾驶行为因子的条件下，30 岁驾驶人的索赔频率为 29.43%（对应于第 300 辆汽车），40 岁驾驶人的索赔频率为 16.84%（对应于第 1100 辆汽车）；在 40 ~ 50 岁区间，随着驾驶人年龄的增加，索赔频率略有上升；在 50 ~ 60 岁区间，索赔频率保持相对稳定；在 60 岁以上的区间，索赔频率随着驾驶人年龄的增加会显著上升，但由于这个区间的车辆数较少，所以结果的可靠性也较低。

图 8 表明，驾驶行为因子对索赔频率的影响也是非线性的，在驾驶行为因子小于 -10 的区间，索赔频率随着驾驶行为因子的增加而上升，但由于这个区间的车辆数较少，估计结果的可靠性较低；在驾驶行为因子大于 -10 的区间，索赔频率随着驾驶行为因子的增加而下降，譬如，第 300 辆汽车的驾驶行为因子小于第 1100 辆汽车，所以其索赔频率大于第 1100 辆汽车。

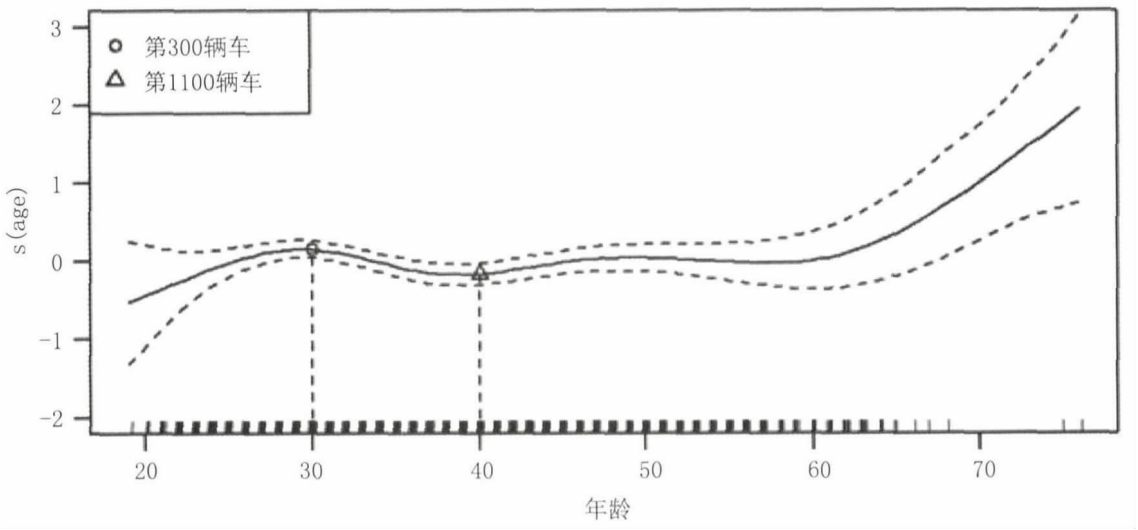

图 7　驾驶人年龄的效应

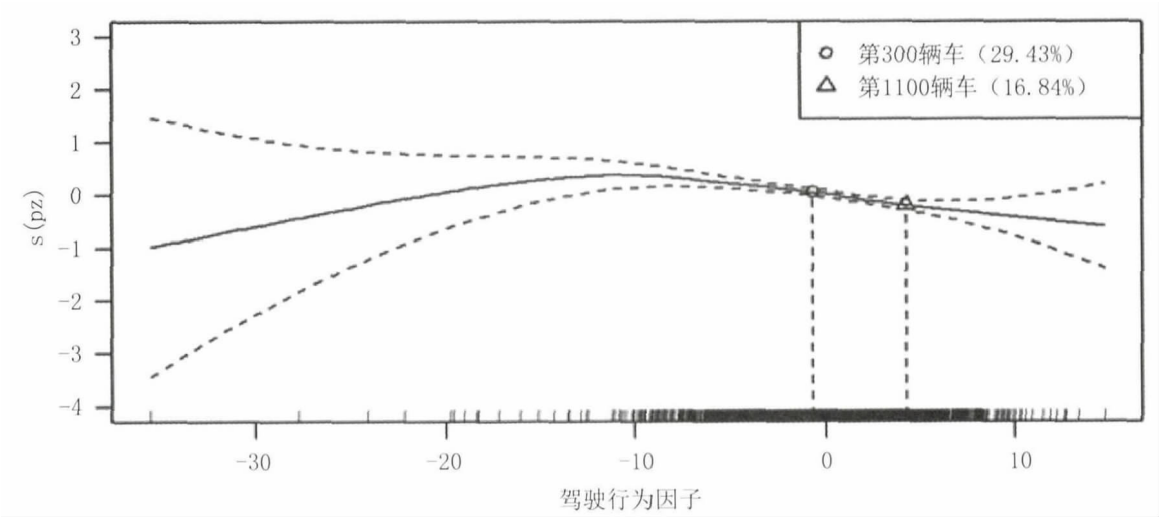

图 8　驾驶行为因子的效应

前面讨论了如何基于现有业务数据提取驾驶行为因子。对于新投保人,保险公司只有该投保人的驾驶行为数据,但在前述的矩阵 Z^0 中并没有与其对应的相对核密度估计,Z^0V_2 中也没有与其对应的驾驶行为因子。此时,为了预测新投保人的索赔频率,可以首先计算该投保人的速度 - 加速度的相对核密度估计 z_{n+1},然后应用式(4)对该向量进行标准化,即令 $z^0_{n+1}=(z_{n+1}-\mu)/\sigma$,最后再计算点积 $pz^2_{n+1}=z^0_{n+1}\cdot V_2$,即可得到新投保人的驾驶行为因子。在已知驾驶人年龄和驾驶行为因子的条件下,应用模型 7 就可以预测其索赔频率。

六、结　论

车联网数据中包含着非常重要的驾驶行为信息,应用这些信息对汽车保险进行定价,有助于提高汽车保险费率的合理性和准确性,也有助于促进整个社会的交通安全。本文基于车联网记录的速度 - 加速度数据,应用核密度估计和主成分分析,提取了一个可以应用于车险费率厘定的驾驶行为因子。实证研究结果表明,该因子的取值与索赔频率之间存在着十分显著的非线性相关关系。

在车联网数据中,除了速度 - 加速度信息外,还包括其他信息,如车辆的横向加速度、GPS 方向和引擎转速等。本文从速度 - 加速度数据中提取了驾驶行为因子,在下一步的研究中,还可以考虑能否从速度 - 横向加速度、速度 - 转弯速度、速度 - 纵向加速度 - 横向加速度等数据中提取驾驶行为因子。此外,本文使用奇异值分解的方法对核密度估计进行降维,但这种方法仅局限于线性变换,在今后的研究中还可以尝试其他非线性的降维方法,如瓶颈神经网络等。

总之,从速度 - 加速度中提取的驾驶行为因子对被保险车辆的索赔频率具有十分显著的非线性影响,在汽车保险中可以作为一个新的费率因子使用,这将有助于提高汽车保险定价结果的合理性和准确性。

[参考文献]

[1] 孟生旺. 非寿险定价[M]. 北京: 中国财政经济出版社,2011.

[2] 孟生旺,李天博,高光远. 基于机器学习算法的车险索赔概率与累积赔款预测[J]. 保险研究,2017,(10):42 - 53.

[3] 孟生旺,王选鹤. GAMLSS 模型及其在车损险费率厘定中的应用[J]. 数理统计与管理,2014,33(04):583 - 591.

[4] 张连增,段白鸽. 行驶里程数对车险净保费的影响研究——基于公路里程对交通事故损失的影响视角[J]. 保险研究,2012,(06):29 - 38.

[5] Ayuso M, Guillen M, Pérez-Marín AM. Telematics and gender discrimination: some usage-based evidence on whether men's risk of accidents differs from women's [J]. Risk, 2016, 4 (2), 10; doi: 10. 3390/risks4020010.

[6] De Jong P, Heller G Z. Generalized linear models for insurance data [M]. Cambridge: Cambridge University Press, 2008.

[7] Gao G, Wüthrich M V. Feature extraction from telematics car driving heatmaps. 2017, SSRN Manuscript ID 3070069.

[8] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. datamining, inference and prediction. 2nd edition [M]. Springer Series in Statistics. 2009.

[9] Leo G. Gradient boosting trees for auto insurance loss cost modeling and prediction [J]. Expert Systems with Applications, 2012, 39(3): 3659 - 3667.

[10] Liu Y, WangJ B, Lv S G. Using multi-class adaboost tree for prediction frequency of auto insurance [J]. Journal of Applied Finance and Banking, 2014, 4(5): 45 - 53.

[11] Verbelen R, Antonio K, ClaeskensG. Unraveling the predictive power of telematics data in car insurance pricing. 2016, SSRN Manuscript ID 2872112.
[12] Weidner W, Transchel F WG, Weidner R. Classification of scale-sensitive telematic observables for riskindividualpricing [J]. European Actuarial Journal, 2016, 6(1): 3 - 24.
[13] Weidner W, Transchel F W G, Weidner R. Telematic driving profile classification in car insurance pricing [J]. Annals of Actuarial Science. 2016, 11(2): 213 - 236.
[14] Wood S N. Generalized additive models: an introduction with R [M]. Chapman & Hall/CRC, 2006.
[15] Wüthrich M V. Covariate selection from telematics car driving data. European Actuarial Journal, 2017, 7(1): 89 - 108.
[16] Wüthrich M V, Buser C. Data analytics for non-life insurance pricing. 2016, SSRN Manuscript ID 2870308.

An Analysis of Rating Factors in Car Insurance Based On Telematics Car Driving Data

GAO Guangyuan[1], MENG Shengwang[1,2]
(1. Center for Applied Statistics, Renmin University of China, Beijing 100872;
2. School of Statistics, Lanzhou University of Finance and Economics, Gansu Lanzhou730020)

Abstract: With the development of vehicle telematics, the value of telematics data becomes apparent in insurance. Telematics car driving data contains detailed driving behavior information, which may be used to improve the traditional rate making models in car insurance. How to extract useful information from telematics data is an on-going research topic and it requires sophisticated statistical methods. This paper analyzed the speed-acceleration data from the telematics car driving data. Kernel density estimation and principal components analysis were applied to extract a driving style index from the speed-acceleration data, and a Poisson generalized additive model for claims frequencies was established. The empirical analysis showed that there was a quite significant non-linear relationship between the extracted driving style index and the claims frequencies. This driving style index can be used as a new rating factor, which will improve the predictive power of the rate making model and lead to a more accurate and reasonable rate.

Key words: telematics; big data; car insurance; rating factor; driving style; claims frequencies

[编辑: 刘延辉]